

TUNING OF MODULATION SPECTRUM DISPERSION PARAMETERS FOR VOICE PATHOLOGY DETECTION

L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente

Universidad Politécnica de Madrid, Madrid, Spain

laureano.moro@upm.es, jorge.gomez.garcia@upm.es, igodino@ics.upm.es

Abstract: Acoustic parameters are frequently used to assess the presence of pathologies in human voice. Many of them have demonstrated to be useful but in some cases its results could be optimized by selecting appropriate working margins. In this study two indices, CIL and RALA, obtained from Modulation Spectra are described and tuned using different frame lengths and frequency ranges to maximize AUC in normal to pathological voice detection. After the tuning process, AUC reaches 0.96 and 0.95 values for CIL and RALA respectively representing an improvement of 16 % and 12 % at each case respect to the typical tuning based only on frame length selection.

Keywords: Modulation Spectrum, voice pathology, CIL, RALA, AUC.

I. INTRODUCTION

The acoustic analysis of voice [1] is a widely used method to assess the presence of a voice pathology due to it is a non-invasive, cost-efficient and easy-to-use technique. Although there are many indices helping clinicians to evaluate the voice perturbations, new parameters are needed to be employed whether in acoustic analysis or as the basis of automatic detectors being used as diagnostic support tools. Moreover, these parameters can be valuable in perceptual assessments to help specialist to increase reliability [2].

Modulation Spectrum (MS) [3] of acoustic signals contains information about the energy relative to modulation frequencies and it can be used as a source of features destined to measure perturbations in voice signal. Many works have used it in the automatic assessing and detection of voice pathologies such as [4] but there is still room for improvement. In this study two measures coming from the histogram of MS are presented. The main objectives of this work are to introduce these new parameters and to determine the optimal operational points for which these can be of use to distinguish between normal and pathological voices, following a simplified version of the methodology used in [5]. In the present case, the tuning

is accomplished considering different frame lengths, acoustic and modulation frequency boundaries in order to obtain optimal Area Under the Curve (AUC) from the Relative Operating Characteristic (ROC) curve and its Standard Error (SE) as suggested in [6] in normal-pathological voice detection.

II. METHODS

A. Modulation Spectrum Dispersion Parameters.

MS provides information about the energy at modulation frequencies that can be found in the carriers of a signal. It is a three-dimensional representation where abscissa usually represents modulation frequency, ordinate axis depicts acoustic frequency and applicate, acoustic energy. To obtain MS, signal passes through a short-Time Fourier Transform (sTFT) filter bank whose output is used to detect amplitude and envelope. This output is finally analyzed using FFT producing a $M \times N$ complex matrix, being M the number of acoustic bands and N the number of modulation bands. Hence, a large amount of data is obtained depending on the size of M and N but in most of the cases MS matrix must be compressed to more specific parameters.

MS allows observing different voice features simultaneously such as fundamental frequency and harmonics and its corresponding modulations. For instance, the presence of tremor, understood as low frequency perturbations of the fundamental frequency, can be easily noticeable since it implies a modulation of pitch as a usual effect of laryngeal muscles improper activity. Fig. 1 shows the MS of a sinusoid without and with amplitude modulation. In the first case (a), only one point stands out from the overall matrix which is located at the central modulation band, corresponding with 0 Hz in the modulation frequency axe. On the other hand, when any type of modulation exists, new emerging points or areas appear in the modulation regions as it is illustrated on Fig. 1 (b). The study of statistics related with these standing out areas can provide new parameters destined to measure voice perturbations.

The two proposed in this study are *Cumulative Intersection Level* (CIL) and *Ratio of points Above Linear Average* (RALA) which are intended to measure dispersion of energy across the modulation frequency axe respect to the acoustic axe. When modulation appears in human voice, due to voluntary or involuntary causes, the energy present in acoustic frequency spreads, going from the acoustic axe to the modulation bands. In these cases, dispersion arises.

Throughout this work, the MS has been calculated using the Modulation Toolbox library ver 2.1 [7].

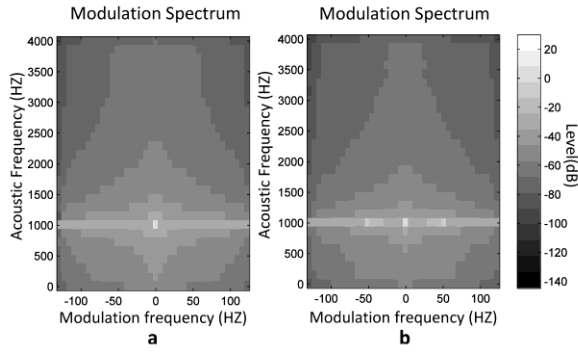


Fig. 1. *Modulation Spectra of 1kHz sinusoid without modulation (a) and with 50 Hz amplitude modulation (b).*

The first parameter, CIL, represents the intersection between the MS histogram increasing and decreasing cumulative curves. Histogram is obtained from MS modulus in logarithmic units (dB) using 29 bins. The higher the number of points in a bin, the nearer CIL index will be to this bin. As it is shown in Fig. 2, CIL tends to be higher in pathological than in healthy voices. On the other hand, RALA is the ratio between points in MS which are over the average of the modulus and the number of points which are above this average. Fig. 3 represents these points in a healthy and a pathological voice. It is noticeable that, as expected, the MS of dysphonic voices present more points above the modulus average.

After describing these new parameters, it is possible to perceive that the MS in Fig. 1 (a) will likely have fewer points over a certain threshold, which can be linear average, than the second one (b). Likewise, the high level bins of the histogram of the second MS (b) will have more cases than those of the first one (a) what will produce a higher CIL.

B. Database

The MEEI voice database is used on this study [8]. From the original 710 recordings, a corpus of 226 including the sustained vowel /ah:/ is selected according to the criteria found in [9]. All the files are

sampled at 25 kHz and 16 bits. Before parameterization, all the recordings are normalized. Voice recordings of normal voices (53 files) have an average duration of 3 s while pathological voices recordings (173 files) have an average duration of 1 s.

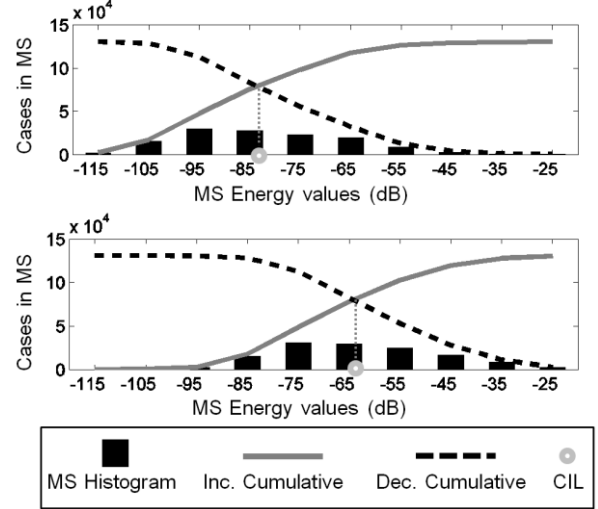


Fig. 2. *CIL calculation in a normal voice (top) and a pathological voice (bottom) diagnosed of bilateral laryngeal tuberculosis. In the second case, histogram presents more points with high levels.*

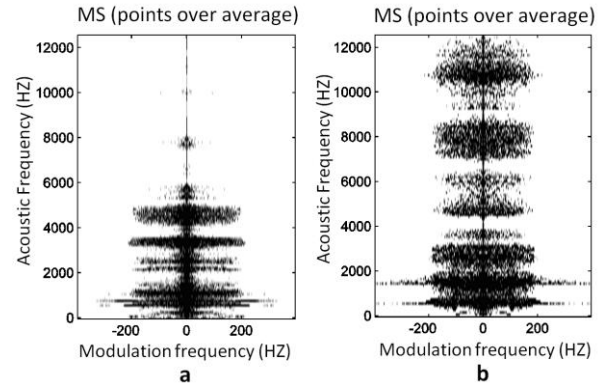


Fig. 3. *Points above (black) and below (white) modulus average in MS for a normal voice (a) RALA = 0.12, and a pathological voice due to bilateral laryngeal tuberculosis (b) RALA = 0.27.*

C. Tuning

To tune CIL and RALA in the voice pathology detection task, three degrees of freedom are selected: frame length, acoustic frequency range and modulation frequency range. Therefore, the corpus is parameterized three times, in which only one degree of freedom is modified at a time while the other two remain fixed. The purpose is to identify which frame

length and frequency ranges lead to the best AUC when using the proposed MS parameters to detect the existence of voice pathology.

Accordingly, in a first stage the corpus is parameterized with CIL and RALA varying only frame lengths in the range of 20 and 200 ms, 50 % overlapping, with fixed acoustic frequency band [0 - 12 kHz] and fixed modulation frequency band [0 - 220 Hz]. This is a basic tuning, widely used to optimize automatic detection systems. With these parameters AUC is calculated. Frame lengths providing the best AUC results are used to re-calculate both parameters separately in a second stage with a new modulation frequency margin, being minimum frequency fixed to 0 Hz and maximum ranging from 20 to 220 Hz in 20 Hz steps. The best AUC results obtained after these two initial stages serve to select optimum modulation frequency range. Using the best frame length and modulation frequency range, a last round of parameterizations and AUC computations are performed by modifying the lower boundary of acoustic frequency between 0 and 1000 Hz in 100 Hz steps and the maximum between 1.2 and 12 kHz in finer steps at low frequencies (300 Hz) and larger steps at high frequencies (from 1 to 3 kHz). After AUC calculation, optimal acoustic frequency range is obtained.

Lastly, two GMM classification systems, one for each parameter separately, are trained to test the ability of CIL and RALA to detect pathological voices. Validation is carried out using a k-Folds technique (8-folds).

III. RESULTS

Regarding the first stage, best results are obtained in frames of 200 ms as it is observed in Fig. 4 (a). Using this frame length, new results are achieved when the maximum modulation frequency is varied, resulting 140 Hz as the optimum operating value as it can be deduced from Fig. 4 (b). Using these settings, a last round of parameterizations is performed varying acoustic frequency ranges, obtaining maximum AUC values of 0.96 for CIL and 0.95 for RALA respectively with SE under 0.01 in both cases.

As it is shown in Fig. 5, the optimum acoustic frequency margin is [0.0 - 1.8 kHz] for CIL and [0.9 - 3.0 kHz] for RALA.

Using these configurations, two GMM systems have been trained following an 8-fold validation scheme in EER. These systems are trained and tested using the tuned parameters, employing the frame length and frequency margins generating the highest AUC. The obtained efficiencies are $92.04 (\pm 3.53) \%$

for CIL and $88.50 (\pm 4.16) \%$ for RALA. DET curves are depicted on Fig. 6.

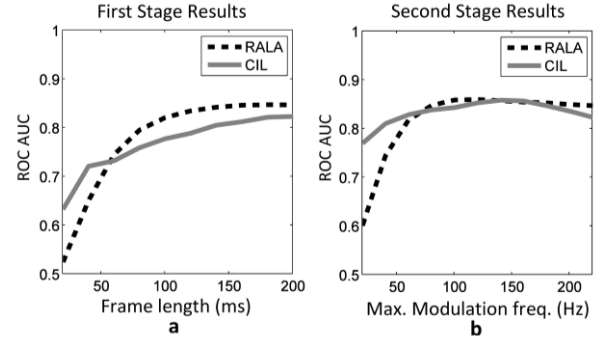


Fig. 4. AUC variation in respect to frame length with fixed acoustic and modulation frequency (a) and to maximum modulation frequency when frame length is 200 ms (b).

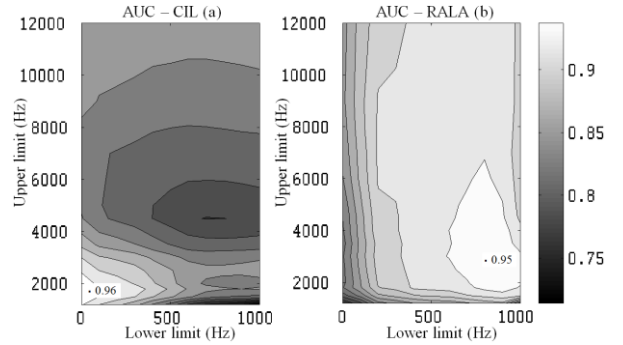


Fig. 5. Influence of acoustic frequency boundaries on ROC AUC for MS parameters CIL (a) and RALA (b).

IV. DISCUSSION

In view of the results it is possible to infer that, at first, there is no relevant information to CIL and RALA over 140 Hz in modulation frequency as the best results are obtained using the margin [0 - 140 Hz] and no improvements were obtained beyond this range. Although this is the optimal margin, Fig. 4. (b) suggests that from 80 Hz as maximum modulation frequency, little improvements are achieved. Hence, it is possible to claim that the most relevant information is contained in the margin [0 - 80 Hz].

Regarding to CIL and the acoustic margins, the most important information seems to be around the fundamental frequency while RALA is optimal above the first formant. Taking into account that CIL aims to indicate if MS has a high number of points or regions with high level respect to the rest of the points, the obtained results suggest that the presence of high-level regions in the acoustic [0.0 - 1.8 kHz] and modulation

[0 – 140 Hz] frequency margins is indicative of the presence of a pathology or a dysfunction in the voice. Comparing the resulting AUC after the third stage with that obtained in the first stage, an improvement of 16 % is achieved. Regarding RALA, the amount of points above average seems to be representative of the presence of a perturbation especially in the range of [0.9 – 3.0 kHz]. In this case, the frequency tuning causes an improvement of 12% in AUC respect to only the frame length tuning.

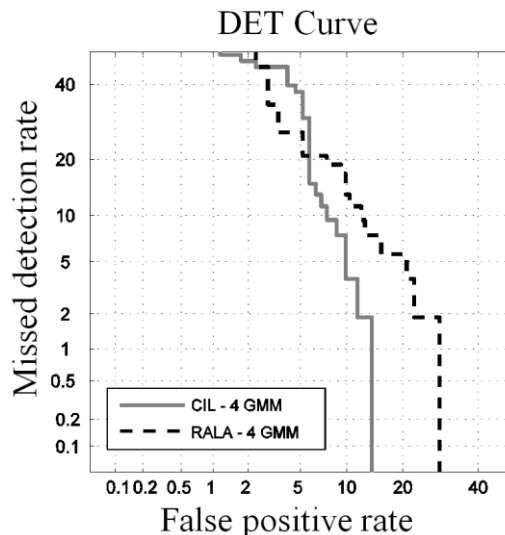


Fig. 6. DET curves for CIL and RALA.

CIL and RALA appear to be suitable for clinical assessment of voice but should be tested using other databases to verify the extent of these results. Moreover, the tuned parameters should be checked in other scenarios such as perceptual assessment of voice quality simulations. Mutual information and correlation can be studied respect to GRBAS subjective assessments.

V. CONCLUSION

In this study two new parameters, CIL and RALA, extracted from MS of voice are presented. These new indices are tuned choosing different frame lengths, acoustic and modulation frequency ranges to obtain maximum AUC values in normal/pathological detection. Results of up to 0.96 of AUC are obtained what suggest that these indices are useful for clinical

applications. Further studies must be performed to evaluate the convenience of these indices and their usefulness in voice quality assessment.

VI. ACKNOWLEDGEMENTS

This research was carried out under grants: TEC201238630-C04-01 from the Spanish Ministry of Education and ayudas para la realización del doctorado (RR01/2011) from Universidad Politécnica de Madrid.

REFERENCES

- [1] C. Sapienza and B. Hoffman Ruddy, *Voice Disorders*. Plural Publishing, 2009.
- [2] I. V. Bele, "Reliability in Perceptual Analysis of Voice Quality," *Journal of Voice*, vol. 19 no. 1, 555-573, 2005
- [3] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Appl. Signal Processing*, 2003.
- [4] J. I. Markaki, M., Stylianou, Y., Arias-Londono, J. D., & Godino-Llorente, "Dysphonia detection based on modulation spectral features and cepstral coefficients," *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 5162–5165, 2010.
- [5] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *Journal of Voice*, vol. 24, no. 1, pp. 47–56, Jan. 2010.
- [6] J. Hanley and B. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases.," *Radiology*, vol. 148, no. 3, pp. 839–843, 1983.
- [7] L. Atlas, P. Clark, and S. Schimmel, "Modulation Toolbox Version 2.1 for MATLAB." University of Washington, 2010.
- [8] "Voice Disorders Database." Massachusetts Eye and Ear Infirmary, Kay Elemetrics Corp., Lincoln Park, NJ., 1994.
- [9] V. Parsa and D. G. Jamieson, "Identification of Pathological Voices Using Glottal Noise Measures," *J Speech Lang Hear Res*, vol. 43, no. 2, pp. 469–485, 2000.